SOVEREIGN: To what extent do different routing mechanisms in sparse MoE models influence inference latency and code gener

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Most R novices will start with Appendix A [A sample session], page 80.This should give some familiarity with the style of R sessions and more importantly some instant feedback on what actually happens.Many users will come to R mainly for its graphical facilities.

1 Introduction

Analysis of: R: A Language and Environment for Statistical Computing. Research goal: To what extent do different routing mechanisms in sparse MoE models influence inference latency and code generation success rates on programming benchmarks like HumanEval and MBPP?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

8 papers retrieved. 2 claims extracted, 2 verified. Tribunal: 7.0/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
Most R novices will start with Appendix A [A sample session], page 80.	✓	0.48
Many users will come to R mainly for its graphical facilities.	✓	0.40

References

- https://doi.org/10.4230/lipics.itp.2023.19
- https://doi.org/10.48550/arxiv.2308.12950
- https://doi.org/10.32614/r.manuals